Module 2

Ananya Verma, Vikas Jadhav

February 2025

Activity 1

Central Limit Theorem

Suppose we are given N independent random variables $\{x_1, x_2 \dots x_N\}$ each of then drawn from some probability distribution $p(x)$. Consider the mean of these random variable

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

which itself is also a random variable. We can then ask the question of what distribution does this random variable, i.e, the mean $(\overline{x})$ follow. According to the central limit theorem, in the limit $N \to \infty$, the means follow a gaussian distribution, regardless of what $p(x)$ is.

To verify the central limit theorem, we sampled several realizations of $\{x_1, x_2 \dots x_N\}$ from a uniform distribution between 1 and 0, for $N = 1, 2, 10$. Then we looked at the distribution of the means of these data points across different realizations. As we can see in Figure 1, for low values of N, the mean follows the same distribution of that the samples follow, which in this case is a uniform distribution. However, as we increase N, we can see that the distribution quickly converges to a gaussian distribution, for values as low as even $N = 10$.

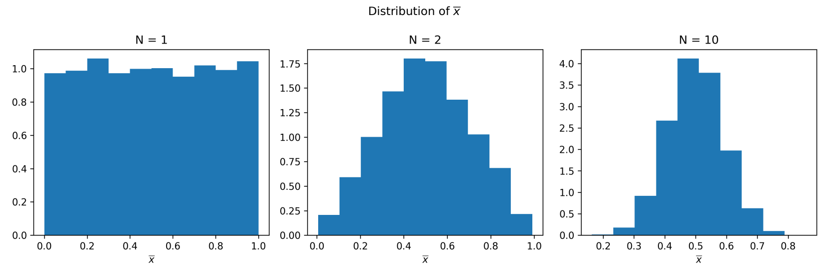

Figure 1: Histogram of the sample means

We can indeed verify that the distribution for $N = 10$ is a Gaussian by

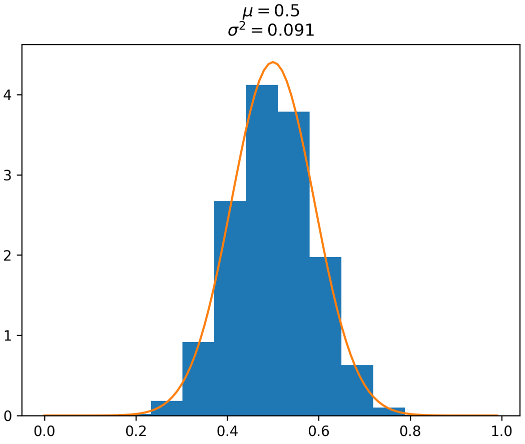

Figure 2: Gaussian fit for histogram of $\overline{x}$ for N=10

fitting the histogram with a Gaussian. The fit obtained is that of a Gaussian with mean = 0.5 and standard deviation = 0.3

Given that any distribution of a random variable leads to a gaussian distribution of the sample means, Gaussian distribution is expected to be ubiquitous in nature. One trivial example would be in statistical physics, where the position of a single particle in the presence of many forces can be modelled as a random walk. In that scenario, the average poition of the particle over time will follow a gaussian distribution.

Activity 2

We generated 1000 samples (x, y) from a 2D Gaussian distribution with $\boldsymbol{\mu} = [0, 0]$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} 4 & 2 \\ 2 & 16 \end{smallmatrix}\right)$

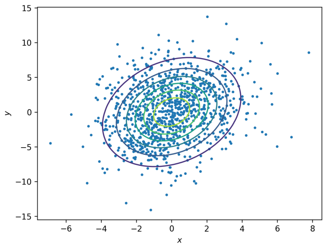

Figure 3: Scatter plot with contour lines

One example would be measurement of position and momentum of harmonic oscillators in canonical ensemble. Since the position and momentum lie on constant energy surfaces that are ellipses, the presence of several energy levels would lead to sampling from different energy contours. Another general example would be any two observables that are linearly related to each other, but also subject to random noise fluctuations that can be modeled using a gaussian distribution, would exhibit such correlation. One such example would be Voltage and Current values in the presence of thermal noise.

Activity 3

We sampled a data set with N points $(x_1, \dots, x_N)$ where each x_i was sampled from a normal distribution with mean, $\mu = 0$, and standard deviation, $\sigma = 0.3$.

The Maximum-Likelihood (ML) estimator for the mean (μ_{ML} and standard deviation (σ_{ML}) for a sample of data is given by,

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2}$$

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{ML})^2 \tag{3}$$

The idea of an estimator is that it is a statistic that aims to capture the property of the true population distribution. We can quantify this statement as follows. Given that θ is the true value of the parameter for the population, and $\hat{\theta}$ is an estimator for that parameter, then we can capture how close it is to the true value by calculating the difference $E[\hat{\theta}] - \theta$, where the expectation value is taken over different realisations of the data.

1. If $E[\hat{\theta}] = \theta$, then we call $\hat{\theta}$ an **unbiased** estimator of θ
2. If $E[\hat{\theta}] \neq \theta$, then we call $\hat{\theta}$ an **biased** estimator of θ

In figure 4, we can see that $E\mu_{ML}$ is close to the true value of $\mu = 0$ for all cases of sample size. Hence, we can conclude that μ_{ML} is a **unbiased** estimator of μ

In figure 5, we can see that for small sample sizes, σ_{ML} varies from the true value significantly and only converges to the true value for large N. Thus we can conclude that for small N, σ_{ML} is a **biased** estimator of σ.
We can correct for this bias by constructing another estimator, $\tilde{\sigma}$, which is obtaining by scaling σ_{ML} by a scale depending on the sample size,

$$\tilde{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_{ML})^2 \tag{4}$$

In Figure 5 we can see that $\tilde{\sigma}$ is a better estimator for σ as it is closer to the true value for small N. However, it can be seen that for large N, both σ_{ML} and $\tilde{\sigma}$ converge to the same value.

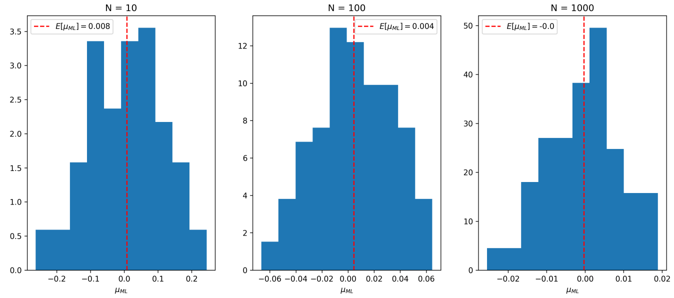

Figure 4: Histogram of values of μ_{ML} for different values of sample size N

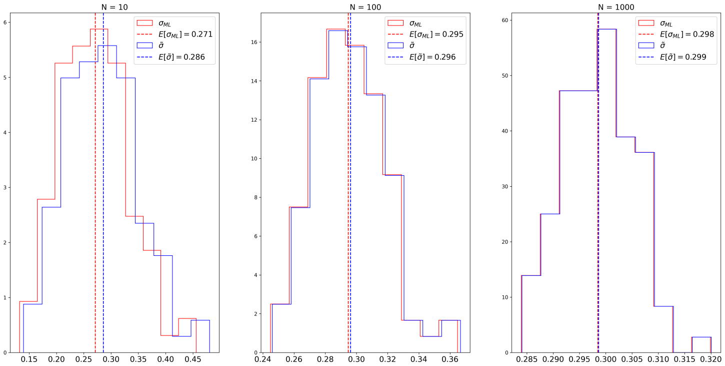

Figure 5: Histograms for the distributions of σ_{ML} (Red) and $\tilde{\sigma}$ (Blue) for different realisations of the data, drawn from the same distribution. The verticals lines represent $E[\sigma_{ML}]$ (Red) and $E[\tilde{\sigma}]$ (Blue)

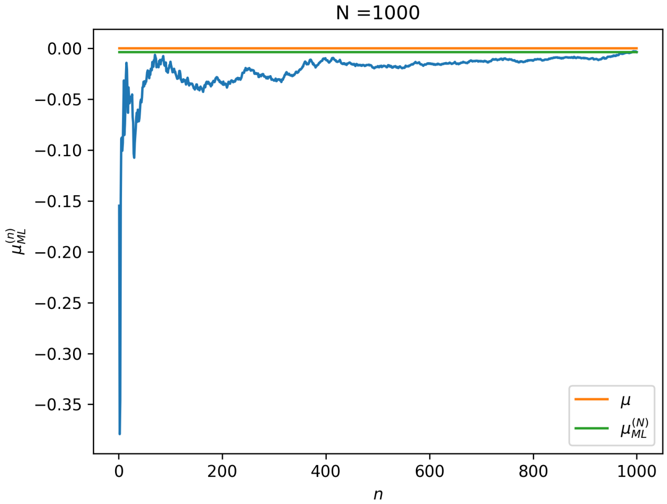

Figure 6: The sequence of mean estimators for a data with N = 1000.

Sequential Estimator

Given a data of N points, $(x_1, \ldots, x_N)$, one can construct ML estimator for the mean using only the first n points.

$$\mu_{ML}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

These estimators $\{\mu_{ML}^{(n)}\}$ form a sequence that eventually converges to the ML estimator considering the whole data, i.e μ_{ML}^{N}. As we have shown earlier, since μ_{ML} is an unbiased estimator of μ, the sequence $\{\mu_{ML}^{(n)}\}$ will also converge to the true value of μ for large N. Figure 6 clearly shows how by including more data points, the sequential estimator converges to the true value.

Activity 4

Conjugate Prior

A conjugate prior is a type of prior distribution that, when combined with a likelihood function from a specific family of distributions, results in a posterior distribution that belongs to the same family as the prior. This makes the mathematical formulation of Bayesian updating simpler and computationally efficient.

Example of Conjugate Priors:

- **For the mean** μ of a Gaussian distribution, a **Normal prior** is conjugate to the **Normal likelihood**. This means the posterior distribution for μ, given a set of data points, is also a Normal distribution.
- **For the variance** σ^2 of a Gaussian distribution, the **Inverse Gamma prior** is conjugate to the Gaussian likelihood. This means the posterior distribution for σ^2 will also be an Inverse Gamma distribution.

Developing the Algorithms for Iterative Bayesian Inference

In Bayesian inference, we iteratively update our belief about the parameters (e.g., μ and σ^2) as more data points are observed. The process of updating is based on **Bayes' Theorem**. When we use conjugate priors, the iterative process becomes simpler because the form of the posterior distribution is the same as the prior distribution.

Steps for Implementing Iterative Bayesian Inference:

1. **Define the Prior:**
   - Choose an appropriate conjugate prior for the parameter of interest (e.g., Normal prior for μ, Inverse Gamma prior for σ^2).
2. **Define the Likelihood:**
   - The likelihood function is defined based on the assumption of the data distribution (e.g., Gaussian likelihood for data generated from a Gaussian distribution).
3. **Use Bayes' Theorem:**
   - Apply Bayes' Theorem to combine the prior and the likelihood to compute the posterior. With conjugate priors, this is done analytically.

     $$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

     where θ represents the parameter(s), D is the data, and $p(D)$ is the evidence (which normalizes the posterior).
4. **Iterative Update:**
   - As new data becomes available, update the posterior to become the new prior for the next iteration.

  - With conjugate priors, the posterior distribution will be in the same form as the prior, but with updated parameters. For example, the posterior for μ is a Normal distribution with updated mean and variance.

5. **Compute the Posterior Mean/Variance:**

   - After each iteration, you can compute the posterior mean (the estimate of the parameter) and/or the posterior variance (the uncertainty in the estimate).

6. **Repeat:**

   - For each new data point, update the prior with the newly calculated posterior and repeat the process until all data is incorporated.

Example Algorithm for μ (Mean of Gaussian Distribution):

1. **Initialize the prior:**

$$p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

2. **For each new data point x_n:**

   - Update the posterior using Bayes' Theorem. With conjugate priors, the posterior for μ is also Normal, with an updated mean and variance:

$$\mu_N = \frac{\frac{1}{\sigma^2}\sum_{n=1}^{N} x_i + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

   - The updated variance is:

$$\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

3. **Repeat:**

   - For every new data point, recalculate the posterior mean μ_N and variance σ_N^2.

Bayesian Interface for mean μ

The Maximum Likelihood Estimate (MLE) for μ is simply the sample mean of the data:

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The Bayesian estimate of μ incorporates prior knowledge, which influences the estimate, especially when the number of observations N is small. As N increases (in our case: $N = 200$, the Bayesian estimate μ_N converges to the MLE μ_{ML}, as the influence of the prior diminishes with more data.

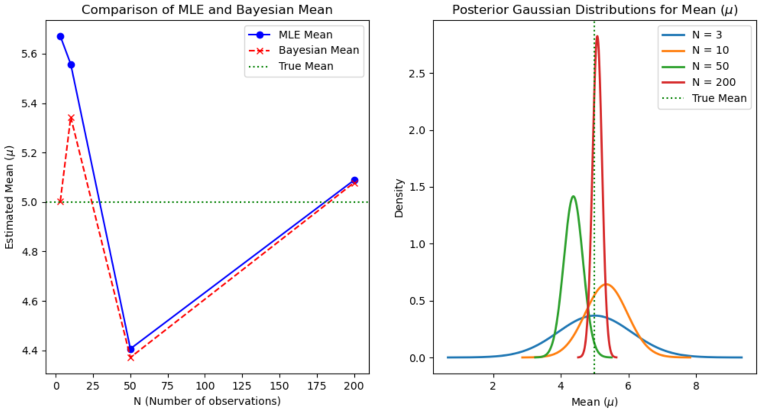

Figure 7: Bayesian Interface of the Mean

Bayesian Interface for variance σ_2

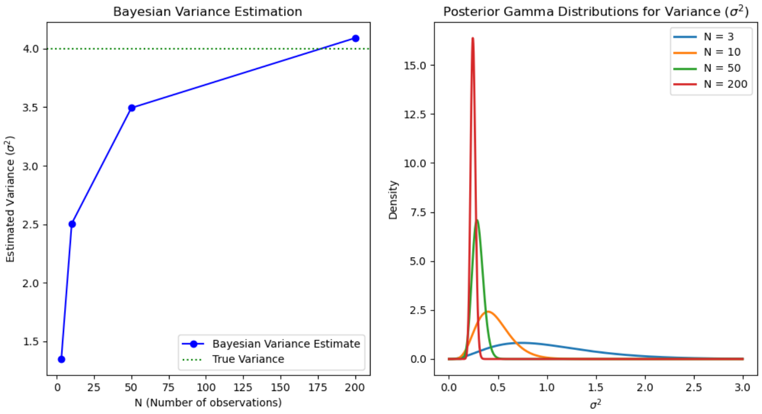

Figure 8: Bayesian Interface of the Variance

Now, we assume that μ is known and need to infer the variance σ^2 using Bayesian methods.

The likelihood function for σ^2 is derived from the Gaussian distribution:

$$p(D \mid \sigma^2) = \prod_{i=1}^{N} N(x_i \mid \mu, \sigma^2)$$

The conjugate prior for the variance σ^2 is the Gamma distribution. This means we assume that:

$$p(\sigma^2) \sim \text{Gamma}(\alpha_0, \beta_0)$$

where α_0 and β_0 are the parameters that define the prior belief about σ^2.

The posterior distribution of σ^2 is also a Gamma distribution, and the updated parameters are:

$$\alpha_N = \alpha_0 + \frac{N}{2}, \quad \beta_N = \beta_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2$$

The Bayesian estimate of σ^2 is given by the posterior mean of the Gamma distribution:

$$\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

This is the expected value of σ^2 given the data and prior.

The Maximum Likelihood Estimate (MLE) for the variance σ^2 is computed as the sample variance:

$$\sigma_{\text{ML}}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

where μ is the known mean.

In contrast, the Bayesian estimate of the variance σ^2 incorporates prior knowledge and is influenced by the conjugate prior. The posterior distribution of σ^2 follows a Gamma distribution, and the Bayesian estimate is given by the posterior mean:

$$\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

As N increases, the Bayesian estimate of σ^2 converges to the MLE, since the influence of the prior diminishes with more data. For small N, however, the Bayesian estimate may differ from the MLE due to the effect of the prior distribution.